CHANGE ANALYSIS

Emanuel Parzen

Department of Statistics

Texas A&M University

Technical Report No. #172

June 1992

Texas A&M Research Foundation

Project No. 6547

'Functional Statistical Data Analysis and Modeling'

Sponsored by the U. S. Army Research Office

Professor Emanuel Parzen, Principal Investigator

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 1992	

4. TITLE AND SUBTITLE
CHANGE ANALYSIS

5. FUNDING NUMBERS
DAAL03-90-G-0069

6. AUTHOR(S)
EMANUEL PARZEN

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Texas A&M University
Department of Statistics
College Station, TX 77843-3143

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARO 27574.9-MA

11. SUPPLEMENTARY NOTES
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
Change Analysis "in the strict sense" is concerned with the problem of detecting and estimating slow and abrupt changes in the probability distributions of successive observations Y(t) of a variable or system. This paper has two goals (1) introduce an approach to Change problems by introducing analysis of Score Change Processes (whose idea is to study if a model to a whole data set fails to fit it by "random walking" the parameter estimating euqations); (2) develop analogies between four basic statistics problems, corresponding to the standard assumptions made about a sequence of observations Y(t), t=1,...,n; test the hypothesis: A: Distribution of specified parametric form, B: Independence, C: Identical distribution, For a sequence of bivariate observations X((t), Y(t)) one would like to test D: Independence of X and Y. Contents are: Introduction, Change Analysis in the strict sense (test Assumption C), Goodness of fit (test Assumption A), Spectral Analysis (test Assumption B), Four phases of change analysis, Parametric scores change analysis, Nonparametric scores change analysis.

14. SUBJECT TERMS
change analysis; changepoint analysis; CUSUMS; parametric change analysis; nonparametric change analysis

15. NUMBER OF PAGES
10

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

CHANGE ANALYSIS

by Emanuel Parzen
Department of Statistics, Texas A&M University
College Station, Texas 77843-3143

Abstract

Change Analysis "in the strict sense" is concerned with the problem of detecting and estimating slow and abrupt changes in the probability distributions of successive observations $Y(t)$ of a variable or system. This paper has two goals (1) introduce an approach to Change problems by introducing analysis of Score Change Processes (whose idea is to study if a model to a whole data set fails to fit it by "random walking" the parameter estimating equations); (2) develop analogies between four basic statistics problems, corresponding to the standard assumptions made about a sequence of observations $Y(t)$, $t = 1, \ldots, n$; test the hypothesis: A: Distribution of specified parametric form, B: Independence, C: Identical distribution, For a sequence of bivariate observations $X((t), Y(t))$ one would like to test D: Independence of X and Y. Contents are: Introduction, Change analysis in the strict sense (test Assumption C), Goodness of fit (test Assumption A), Spectral Analysis (test Assumption B), Four phases of change analysis, Parametric scores change analysis, Nonparametric scores change analysis.

1. Introduction

Data $Y(1), \ldots, Y(n)$ which can be regarded as continuous random variables observed sequentially can be called indexed data or a time series. Classic statistical inference makes three basic assumptions:

Assumption A. Probability law of each Y has probability density belonging to a known parametric family of probability densities $f(y; \theta)$.

Assumption B. Random variables $Y(1), \ldots, Y(n)$ are independent.

Assumption C. Random variables $Y(1), \ldots, Y(n)$ are identically distributed.

Methods for detecting (and estimating) the fit (and the nature of violations) of these assumptions in our opinion can be respectively related to three parallel theories:

Theory C. Changepoint analysis or change analysis (in the strict sense).

Theory B. Spectral analysis (time series analysis in the frquency domain).

Theory A. Goodness of fit.

We believe that one can define a theory, called Comparison Change Analysis, which is intended to study analogies between theories A,B,C (and bring the insights of the theories that are more developed, such as spectral analysis, to less developed ones). General accounts of this theory are given in Parzen (1992), (1991).

The assumption that the data is observed sequentially, which may seem to limit the applicability of Change Analysis, is dropped when the analogies are extended to the bivariate data analysis problem which considers independent bivariate data $(X(t), Y(t))$, $t = 1, \ldots, n$, and desires to model the relation between X and Y and in particular to test

Assumption D. X and Y are independent random variables.

A general non-parametric theory of testing assumption D can be related to

Research supported by the U. S. Army Research Office.

Theory D. Change analysis (random effect).

Analogies between theories A to D are obtained from the facts that in each problem the first step in analysis is to define a dynamic statistic which is a function on the unit interval [0,1] whose asymptotic distribution (under the null hypothesis that the assumptions are true) is either a Brownian Bridge or a related process. The test statistics in each theory are analogous to the nonparametric test statistics that statisticians have developed to test goodness of fit for equality of two distributions.

Textbooks imply it is difficult to choose among the many test statistics for goodness of fit and analogous testing problems; we believe we should be optimistic about our ultimate ability to develop procedures for adaptively choosing appropriate test statistics which not only test the null hypothesis but also suggest likely models instead of only rejecting the null hypothesis.

2. Change analysis in the strict sense (test Assumption C)

The theory of change analysis in the strict sense considers data $Y(t), t = 1, \ldots, n$ which represents a transformation of observed data (the identity transformation leaves the data unchanged).

Let Y^- denote the sample mean (an estimator of the true mean μ if the data are identically distributed). Let $\sigma_Y\hat{}$ denote a suitable estimator (such as the sample standard deviation) of the true standard deviation σ_Y of the data under the assumption of identical distribution (which is assumed to be finite).

The data Y is transformed to normalized data

$$Y\tilde{}(t) = (Y(t) - Y^-)/\sigma_Y\hat{}.$$

We plot the normalized data as a sample *change density* $c\tilde{}(\tau)$, $0 < \tau < 1$, defined to be a piecewise constant function whose value is equal to $Y\tilde{}(j)$ on the interval $(j-1)/n < \tau < j/n$, for $j = 1, \ldots, n$. Note that $\int_0^1 c\tilde{}(\tau)d\tau = 0$, $\int_0^1 c\tilde{}^2(\tau)d\tau = 1$.

CUSUMS (cumulative sums) are becoming increasingly important diagnostic tools to look for patterns in indexed data. They are related to the sample *change process* on $0 < \tau < 1$

$$C\tilde{}(\tau) = \int_0^\tau c\tilde{}(t)dt.$$

The points $\tau = j/n$ for $j = 1, \ldots, n$ are called "exact" values of τ; at these points $C(\tau)$ equals a cumulative sum:

$$C\tilde{}(j/n) = (1/n)\sum_{k=1}^{j} Y\tilde{}(k) = (j/n)Y_j^-.$$

To understand why the change process is an effective means of detecting change in the data consider its behavior under two models for $Y(.)$.

If $Y(.)$ is deterministic and linear, say $Y(t) = t$, then at exact $\tau = j/n$ approximately

$$c\tilde{}(\tau) = Y\tilde{}(j) = 12^{.5}(\tau - .5),$$
$$C\tilde{}(\tau) = (-.5)12^{.5}\tau(1-\tau).$$

The graph of $C\tilde{}(\tau)$ when $Y(.)$ is linear is a parabola that goes from 0 to 0 with mimimum value at $\tau = .5$.

If $Y(.)$ is random (independent identically distributed), the stochastic process $C^{\sim}(\tau)$, $0 < \tau < 1$, can be shown to be asymptotically distributed (as the sample size n tends to infinity) as a Brownian Bridge stochastic process $B(\tau)$, $0 < \tau < 1$, which is a zero mean Gaussian process with covariance kernel $E[B(s)B(t)] = min(s,t) - st$. Note that $B(0) = B(1) = 0$, and

$$\text{Variance}[B(\tau)] = \tau(1-\tau).$$

To test for departures from Assumption C (identical distribution) one tests if the observed change process $C^{\sim}(\tau)$ is significantly different from a sample curve of a Brownian Bridge which can be expected to be a wiggly (non-smooth) curve oscillating about the horizontal axis.

A related process that plays a central role in change analysis is the Change Test Process

$$CT^{\sim}(\tau) = C^{\sim}(\tau)/(\tau(1-\tau))^{.5}.$$

The fundamental role of the change test process starts with the fact that for fixed $\tau = j/n$, $CT^{\sim}(\tau)$ can be shown to be a monotone transformation of the classic two-sample Student's t-test statistic of the null hypothesis $\mu_1 = \mu_2$ in the model $Y(1), \ldots, Y(j)$ is Normal(μ_1, σ^2) and $Y(j+1), \ldots, Y(n)$ is Normal(μ_2, σ^2). The sample means and variances of the two samples $Y(1), \ldots, Y(j)$ and $Y(j+1), \ldots, Y(n)$ are respectively denoted $\mu_1\hat{}, S_1\hat{}^2$ and $\mu_2\hat{}, S_2\hat{}^2$. The pooled sample variance is

$$S\hat{}^2 = \tau S_1\hat{}^2 + (1-\tau)S_2\hat{}^2.$$

One can verify that

$$\hat{\sigma}_Y^2 = S\hat{}^2 + (\tau(1-\tau))(\mu_1\hat{} - \mu_2\hat{})^2,$$
$$\mu_1\hat{} - Y^{\sim} = (1-\tau)(\mu_1\hat{} - \mu_2\hat{})$$

The classic two-sample Student's t-test statistic is $n^{.5}\ T$, defining

$$T = (\tau(1-\tau))^{.5}(\mu_1\hat{} - \mu_2\hat{})/S\hat{}.$$

Define R, a "correlation version" of T, by

$$R^2 = T^2/(1+T^2), T^2 = R^2/(1-R^2).$$

Then

$$R^2 = \tau(1-\tau)(\mu_1\hat{} - \mu_2\hat{})^2/\hat{\sigma}_Y^2$$

and one concludes that $CT^{\sim}(\tau)$ is, like R, a correlation type statistic since

$$R^2 = (\tau/(1-\tau))(\mu_1\hat{} - Y^{\sim})^2/\hat{\sigma}_Y^2$$
$$= |CT^{\sim}(\tau)|^2.$$

We can consequently express Student's t-test statistic T as a monotone function of $CT^{\sim}(\tau)$ since $T = R/(1-R^2)^{.5}$.

Let $\tau\hat{}$ denote the value among the exact values $\tau = j/n$ (for $j = 1, \ldots, n-1$) at which the absolute value of $CT^{\sim}(\tau)$ achieves its maximum. Under the assumption of at most one change in the distribution of $Y(.)$, $CT^{\sim}(\tau\hat{})$ is a test statistic for change

and its time of occurence is consistently estimated by $\tau\hat{}$ (a result established by Carlstein (1988)).

3. Goodness of fit (Test Assumption A)

One of the most extensive and least applied branches of statistical theory is the theory of goodness of fit of probability models to observed data. Despite its importance (both for theory and practice) it appears to be sparsely taught to graduate students in statistics. The chi-squared goodness of fit test introduced by Karl Pearson in 1900 is regarded as one of the top 20 achievements in modern science. How can one explain the neglect of instruction in its theory? One explanation may be that its theory is often taught rigorously as a study in pure probability theory rather than developed vigorously for its statistical interpretation.

Let $Y(t), t = 1, \ldots, n$, be a random sample of a continuous random variable with true distribution $F(y) = F(y; \theta_0)$ belonging to a finite parametric family $F(y; \theta)$. The true quantile function is $F^{-1}(u; \theta_0)$, $0 < u < 1$. The sample distribution function is denoted

$$F\tilde{}(y) = \text{fraction of sample } \leq y.$$

Let $\theta\hat{}$ denote the maximum likelihood estimator of θ. Stochastic processes whose asymptotic properties are of interest (for both theory and practice) are

$$F\tilde{}(y) - F(y; \theta_0),$$
$$F\tilde{}(y) - F(y; \theta\hat{}),$$
$$F(y; \theta\hat{}) - F(y; \theta_0),$$

evaluated at $y = F^{-1}(u; \theta_0)$, $0 < u < 1$. We denote such a process $C\tilde{}(u), 0 < u < 1$, to emphasize its analogy to a sample change process. We use functions of u to study changes of distribution, and functions of τ to study changes of models fitting data.

The testing and estimation procedures of goodness of fit theory can be organized into four phases summarized (in section 5) in our discussion of the four phases of change analysis.

4. Spectral Analysis (Test Assumption B)

One approach to testing the assumption of independence is to consider as an alternative hypothesis for the data $Y(t), t = 1, \ldots, n$, that it is a zero mean stationary time series with covariance function, defined for $v = 0, \pm 1, \pm 2, \ldots,$

$$R(v) = E[Y(t)Y(t - v)]$$

and spectral density function, defined for $0 < \omega < 1$,

$$f(\omega) = \sum_{v=-\infty}^{\infty} R(v) e^{-i2\pi\omega v}$$

The sample spectral density is defined

$$f\tilde{}(\omega) = (2\pi n)^{-1} \left| \sum_{t=1}^{n} Y(t) e^{i2\pi\omega t} \right|^2$$

with sample distribution function (on $0 < \omega < 1$)

$$F^{\sim}(\omega) = \int_0^{\omega} f(\lambda)d\lambda$$

Normalized versions of these functions are

$$f^{*\sim}(\omega) = f^{\sim}(\omega)/F^{\sim}(1),$$
$$F^{*\sim}(\omega) = F^{\sim}(\omega)/F^{\sim}(1).$$

Analogues of the sample change density and sample change process are

$$c^{\sim}(\omega) = f^{*\sim}(w) - 1,$$
$$C^{\sim}(\omega) = F^{*\sim}(w) - w.$$

5. Four phases of change analysis

A sample change process $C^{\sim}(\tau), 0 < \tau < 1$, is a dynamic statistic (sample path of a stochastic process) which often can be shown to satisfy under the null hypothesis of "no change" the null hypothesis $H_0 : C^{\sim}(.)$ is a Brownian Bridge (or a related hypothesis). The statistical analysis of $C^{\sim}(.)$ has four phases:

Phase 1: ***Graphical analysis***; is the plot of $C^{\sim}(\tau)$, $0 < \tau < 1$, oscillatory, a deterministic parabola, other patterns.

Phase 2: ***Non-linear functionals.*** One tests H_0 by computing the values of test statistics (whose asymptotic distributions under H_0 can be deduced from the theory of empirical processes)

$$\int_0^1 |C^{\sim}(\tau)|^2 d\tau,$$
$$\int_0^1 (|C^{\sim}(\tau)|^2/\tau(1-\tau))d\tau,$$
$$\max_{0<\tau<1} |C^{\sim}(\tau)|,$$
$$\max_{\tau=j/n} |C^{\sim}(\tau)|/\tau(1-\tau).$$

Phase 3: ***Linear functionals.*** For various score functions $K(\tau)$, called ***change score*** functions, one computes the linear functional (or component)

$$C^{\sim}(K) = \int_0^1 K(\tau)dC^{\sim}(\tau) = \int_0^1 K(\tau)c^{\sim}(\tau)d\tau$$

One can often write approximately

$$C^{\sim}(K) = (1/n)\sum_{j=1}^{n} K((j-.5)/n)c^{\sim}((j-.5)/n)$$

The score function is usually chosen as a sequence of Orthonormal functions $\psi_1(.), \psi_2(.), \ldots$, especially the Legendre polynomials, which test against patterns in the change density $c^{\sim}(\tau)$.

The key to change analysis is to choose transformations of data (score the data) which are most powerful for detecting change. From the sample change processes, suitable linear functionals (score the change) are formed. These linear functionals are called "double score components". One can define bivariate density functions $d(\tau, u)$, $0 < \tau < 1, 0 < u < 1$, of which double score functions are diagnostics. Choice of data score functions are motivated in sections 6 and 7 parametrically and non-parametrically, respectively.

Phase 4: *Density estimation.* By one of the many methods available in the theory of curve smoothing (kernel methods, splines, exponential methods, wavelets, etc.) form a smooth estimator $\hat{c}(\tau)$ of the change density.

An exposition of the theory of these phases would require a book and is beyond the scope of this paper. Our goal in this paper is to outline the phases and to explain how we choose transformations of the original data from which to form a change process.

6. Parametric scores change analysis

To detect change over time in a sequence one must have some prior opinion about the ways in which the probability distribution of the observations may be changing (such as in location, scale, skewness, etc). Sample change processes are formed for transformed data, where the transformation is called intuitively a *data score* function. The most powerful data transformations are essentially the sufficient statistics, or more precisely the Fisher score functions, when one has a parametric model $f(y; \theta)$ for a random sample $Y(t), t = 1, \ldots, n$, where $\theta = (\theta_1, \ldots, \theta_k)$.

The maximum likelihood estimator $\theta\hat{}$ is obtained by maximizing the average log-likelihood

$$L(\theta) = (1/n) \sum_{t=1}^{n} \log f(Y(t); \theta)$$

Define score functions

$$S_j(Y; \theta) = \partial / \partial \theta_j \log f(Y; \theta)$$

The maximum likelihood estimator is the solution of the *estimating equations* for $j = 1, \ldots, k$

$$(1/n) \sum_{t=1}^{n} S_j(Y(t); \theta\hat{}) = 0.$$

Our approach to change analysis asks if for every potential changepoint $\tau = m/n$ the parametric model with $\theta = \theta\hat{}$ fits the data $Y(t)$, $t = 1, \ldots, m$, up to the time m in the sense that approximately

$$(1/n) \sum_{t=1}^{m} S_j(Y(t); \theta\hat{}) = 0.$$

We define the *score change* process to linearly interpolate its values at $\tau = m/n$, for $m = 1, \ldots, n$

$$C\tilde{}(\tau; S_j) = (1/n) \sum_{t=1}^{m} S_j^*(Y(t); \theta\hat{})$$

where

$$S_j^*(Y; \theta\hat{}). = S_j(Y; \theta\hat{}) / E_{\theta\hat{}}^5[S_j(Y; \theta\hat{})].$$

We form k score change processes, for $j = 1, \ldots, k$.

We call this approach "random walk your normalized scores." We are developing the probability theory of the score change processes.

These theoretical concepts can best be understood through examples. Consider a gamma distribution model

$$f(y; \nu, \theta) = (\theta^{\nu}\Gamma(\nu))^{-1} x^{\nu-1} \exp(-y/\theta)$$

where θ is a positive scale parameter, assumed unknown, and ν is a positive shape parameter, assmed known. One can show that the score function of the parameter θ is

$$S(Y; \theta) = (1/\theta)((Y/\theta) - \nu);$$

the maximum likelihood estimator is

$$\hat{\theta} = Y^{-}/\nu;$$

the normalized score function evaluated at the maximum likelihood estimator of the parameter may be shown to be

$$S^{*}(Y(t); \hat{\theta}) = \nu^{.5}((Y(t)/Y^{-}) - 1).$$

To test the observations $Y(.)$ for change, one forms the maximum likelihood score change process $C^{\sim}(\tau; S^{*})$, $0 < \tau < 1$, and tests if this dynamic statistic is significantly different from a sample path of a Brownian Bridge stochastic process. A linear functional of the change process corresponding to the score function

$$K(\tau) = 12^{.5}(\tau - .5)$$

is

$$C^{\sim}(K, S^{*}) = (1/n)\sum_{t=1}^{n}(12\nu)^{.5}((Y(t)/Y^{-}) - 1)((t - .5)/n)$$

$$= (12\nu)^{.5}(1/n)\sum_{t=1}^{n} Y(t)((t - .5)/n)/Y^{-}$$

Under the null hypothesis of no change the asymptotic distribution of $n^{.5}C^{\sim}(K, S^{*})$ is Normal(0,1).

An example of an application of this statistic is in Hsu (1979) where it is presented as a test designed for a small change in the scale parameter θ of an independent Gamma distributed sequence, derived by Kander and Zacks (1966) by a Bayesian analysis assuming the changepoint τ is uniformly distributed in time. This test statistic is derived in our approach as analogous to a component in standard goodness of fit analysis.

7. Nonparametric scores change analysis

Our approach to change analysis recommends that one compute and interpret several change processes formed from several transformations of the original data. In addition to (or instead of) various parametric score change processes, one can define various nonparametric score processes for a data sequence $Y(t)$, $t = 1, \ldots, n$. Define:

sample distribution function $\tilde{F}(y)$;

sample probability mass function $\tilde{p}(y)$ =fraction of sample equal to y;

mid-distribution function $\tilde{P}(y) = \tilde{F}(y) - .5\tilde{p}(y)$.

The mid-rank data transformation forms $\tilde{P}(Y(t))$, $t = 1, \ldots, n$. When all Y values are distinct, $\tilde{P}(Y(t)) = (\text{Rank}(Y(t)) - .5)/n$; we recommend this definition of mid-ranks over the most used definition Rank $(Y(t))/(n+1)$.

One chooses a data score function $J(u)$, $0 < u < 1$, suitable for testing nonparametrically various types of changes in the distribution of the data (especially changes in location or scale parameters). A typical choice for $J(u)$ is a Legendre polynomial normalized to satisfy

$$\int_0^1 J(u)du = 0, \int_0^1 J^2(u)du = 1.$$

Apply the four phases of change analysis to the transformed data sequence $J(\tilde{P}(Y(t))$. In the third phase one examines and interprets linear functional tests for change of the form

$$\tilde{C}(K, J) = (1/n)\sum_{t=1}^{n} K((t - .5)/n)J(\tilde{P}(Y(t))$$

for suitable change score functions $K(\tau)$. One can usually show that under the null hypothesis of no change the asymptotic distribution of $n^{.5}\tilde{C}(K, J)$ is Normal(0,1).

References

Carlstein, E. (1988). "Nonparametric change-point estimation, *The Annals of Statistics*, **16**, 188-197.

Hsu, D. A. (1979) "Detecting shifts of parameter in gamma sequences with applications to stock price and air traffice flow analysis," *Journal of the American Statistical Association*, **74**, 31-40.

Kander, Z., and Zacks, S. (1966). "Test procedures for possible changes in parameters of statistical distributions occurring at unknown time points," *Annals of Mathematical Statistics*, **37**, 1196-1210.

Parzen, Emanuel. (1991). "Unification of statistical methods for continuous and discrete data," *Proceedings Computer Science-Statistics INTERFACE '90*, (eds. C. Page and R. LePage), Springer Verlag: New York, 235-242.

Parzen, Emanuel. (1992). "Comparison change analysis." *Nonparametric Statistics Internatinal Symposium*, (ed. A. K. Saleh), Elsevier: Amsterdam.